Python程式設計 期末報告

主題：網路爬蟲

台灣大學 經濟三 楊智宇

主要功能：

1. 可自行輸入欲爬蟲的ptt網址，及欲抓取的頁數。
2. 印出抓取的結果，並將輸出格式化。
3. 可選擇是否要匯出成csv檔。

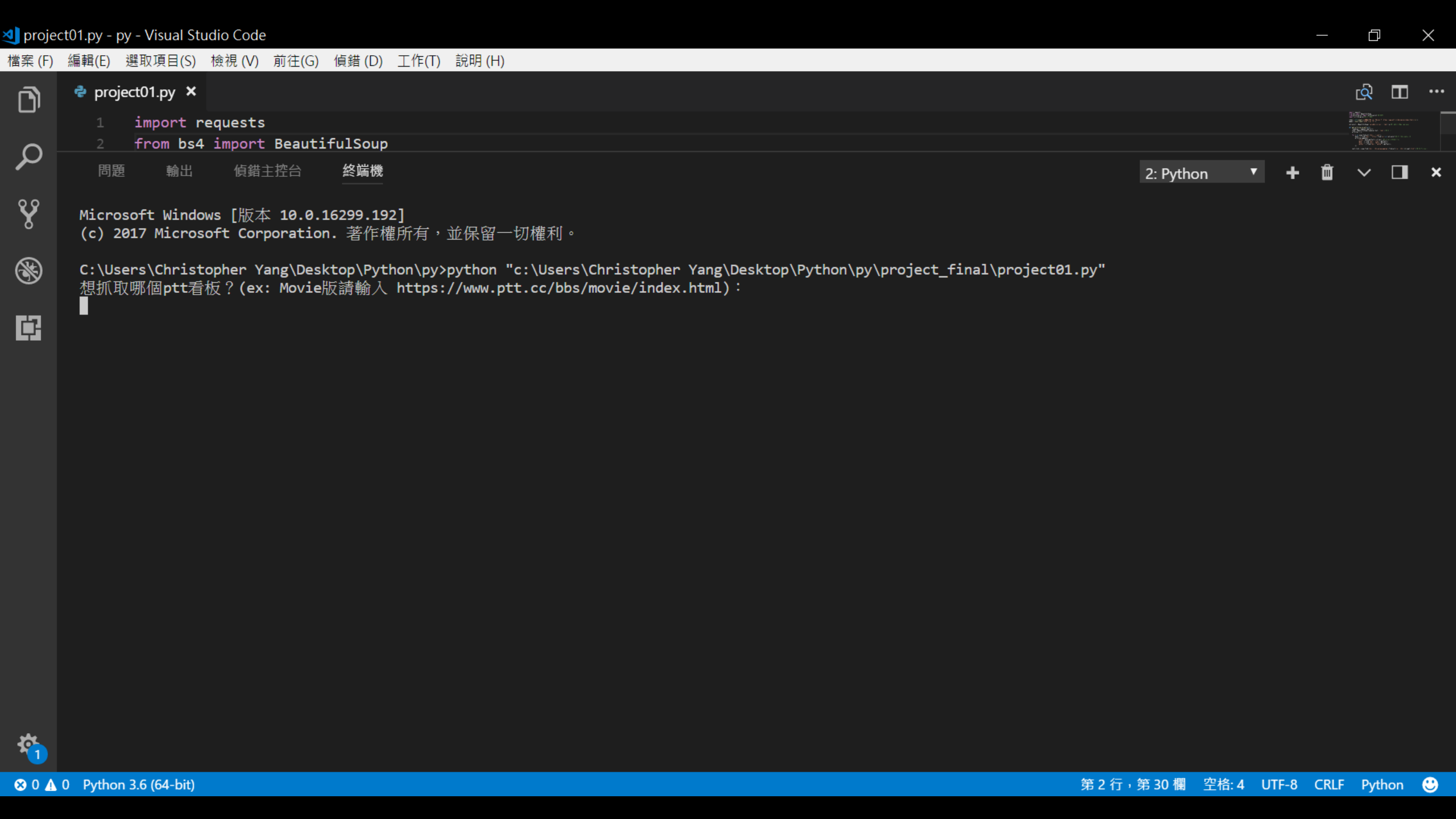
https://goo.gl/ivMCMQ

程式碼

```python
import requests
from bs4 import BeautifulSoup
from pretty_print import pretty_print ## 美化輸出
import urllib.parse ## url 相關應用

index = str(input('想抓取哪個ptt看板？(ex: Movie版請輸入 https://www.ptt.cc/bbs/movie/index.html)：\n'))
pages = eval(input('想抓取幾頁呢？ex: 5：'))

not_exist = BeautifulSoup('<a>(本文已被刪除)</a>', 'lxml').a ## '本文已被刪除'的結構不同，自行生成<a>

def get_articles_on_ptt(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'lxml') ## 網頁原始碼
    articles = []

    for i in soup.find_all('div', 'r-ent'):
        meta = i.find('div', 'title').find('a') or not_exist ## 當爬到是空的，讓 meta = 0
        articles.append({
            'title': meta.getText().strip(), ## strip 去除頭尾字符，預設是空白
            'push': i.find('div', 'nrec').getText(),
            'date': i.find('div', 'date').getText(),
            'author': i.find('div', 'author').getText(),
        })

    next_link = soup.find('div', 'btn-group-paging').find_all('a', 'btn')[1].get('href') ## 控制頁面選項(上一頁)

    return articles, next_link

def get_pages(num): ## 要爬幾頁
    page_url = index
    all_articles = []
```

```python
def get_pages(num): ## 要爬幾頁
    page_url = index
    all_articles = []

    for j in range(num):
        articles, next_link = get_articles_on_ptt(page_url)
        all_articles += articles
        page_url = urllib.parse.urljoin(index, next_link) ## 將上一頁按鈕的網址和 index 網址比對後取代

    return all_articles

data = get_pages(pages)

for k in data:
    pretty_print(k['push'], k['title'], k['date'], k['author'])

csv_or_not = input('輸入 y 以匯出成csv檔，輸入其他結束程式：')

if csv_or_not == 'y':
    board = index.split('/')[-2]
    csv = open('./ptt_%s版_前%d頁.csv'%(board, pages), 'a+', encoding='utf-8')
    csv.write('推文數,標題,發文日期,作者ID,\n')
    for l in data:
        l['title'] = l['title'].replace(',', '，') ## 與用來分隔的逗點作區別
        csv.write(l['push'] + ',' + l['title'] + ',' + l['date'] + ',' + l['author'] + ',\n')
    csv.close()
    print('csv檔案已儲存在您的資料夾中。')
else:
    quit()
```

執行過程


project01.py - py - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)
project01.py
1 import requests
2 from bs4 import BeautifulSoup
問題 輸出 偵錯主控台 終端機
2: Python
Microsoft Windows [版本 10.0.16299.192]
(c) 2017 Microsoft Corporation. 著作權所有，並保留一切權利。
C:\Users\Christopher Yang\Desktop\Python\py>python "c:\Users\Christopher Yang\Desktop\Python\py\project_final\project01.py"
想抓取哪個ptt看板？(ex: Movie版請輸入 https://www.ptt.cc/bbs/movie/index.html)：
0 0 Python 3.6 (64-bit)
第 2 行，第 30 欄 空格: 4 UTF-8 CRLF Python

project01.py - py - Visual Studio Code

檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)

project01.py

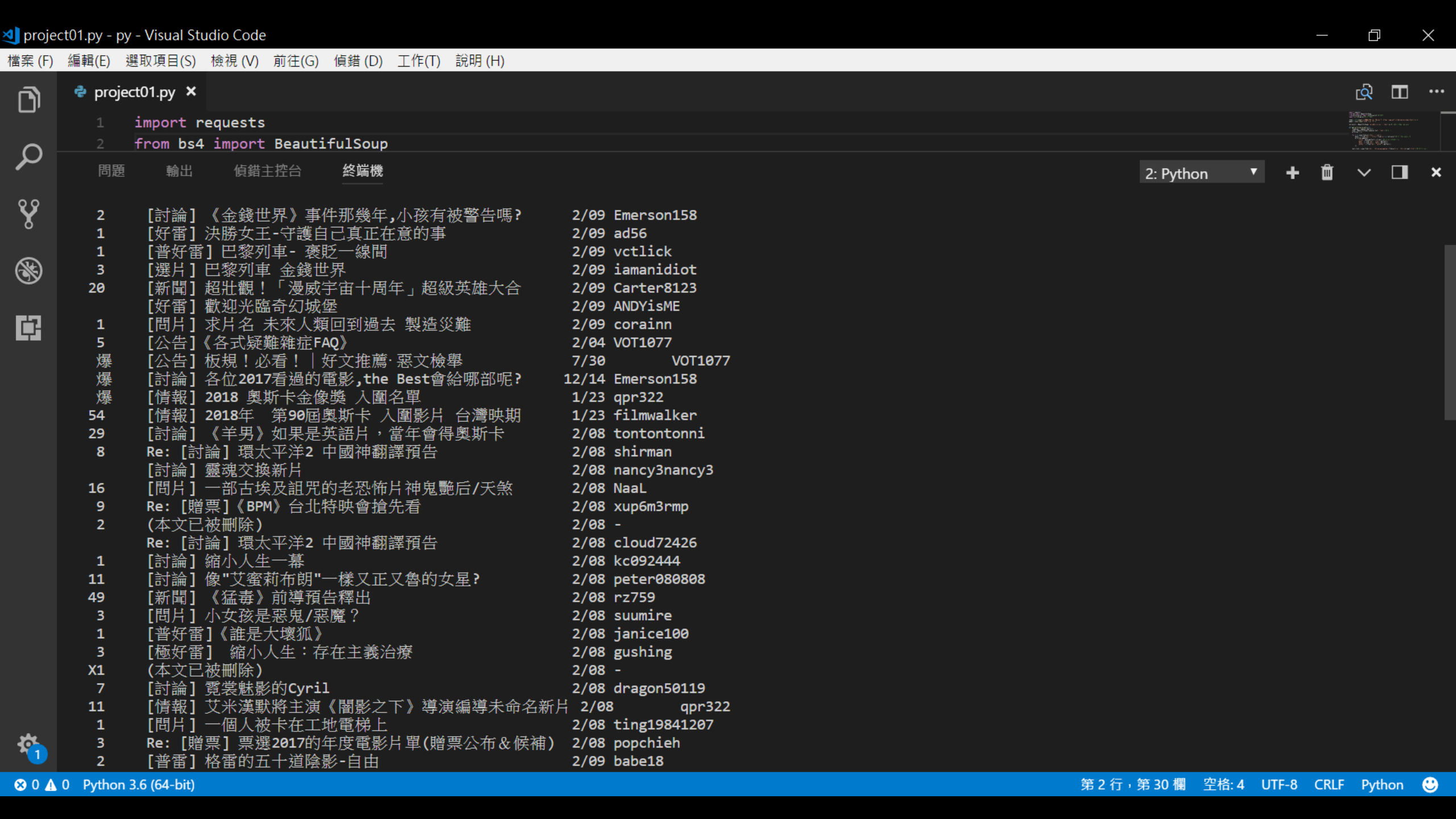
```
1  import requests
2  from bs4 import BeautifulSoup
```

問題 輸出 偵錯主控台 終端機

2: Python

```
Microsoft Windows [版本 10.0.16299.192]
(c) 2017 Microsoft Corporation. 著作權所有，並保留一切權利。

C:\Users\Christopher Yang\Desktop\Python\py>python "c:\Users\Christopher Yang\Desktop\Python\py\project_final\project01.py"
想抓取哪個ptt看板？(ex: Movie版請輸入 https://www.ptt.cc/bbs/movie/index.html)：
https://www.ptt.cc/bbs/movie/index.html
想抓取幾頁呢？ex: 5：
```

0 0 Python 3.6 (64-bit) 第 2 行，第 30 欄 空格: 4 UTF-8 CRLF Python


project01.py - py - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)
project01.py
1 import requests
2 from bs4 import BeautifulSoup
問題 輸出 偵錯主控台 終端機
2: Python
2 [討論]《金錢世界》事件那幾年,小孩有被警告嗎? 2/09 Emerson158
1 [好雷]決勝女王-守護自己真正在意的事 2/09 ad56
1 [普好雷]巴黎列車- 褒貶一線間 2/09 vctlick
3 [選片]巴黎列車 金錢世界 2/09 iamanidiot
20 [新聞]超壯觀！「漫威宇宙十周年」超級英雄大合 2/09 Carter8123
[好雷]歡迎光臨奇幻城堡 2/09 ANDYisME
1 [問片]求片名 未來人類回到過去 製造災難 2/09 corainn
5 [公告]《各式疑難雜症FAQ》 2/04 VOT1077
爆 [公告]板規！必看！｜好文推薦·惡文檢舉 7/30 VOT1077
爆 [討論]各位2017看過的電影,the Best會給哪部呢? 12/14 Emerson158
爆 [情報]2018 奧斯卡金像獎 入圍名單 1/23 qpr322
54 [情報]2018年 第90屆奧斯卡 入圍影片 台灣映期 1/23 filmwalker
29 [討論]《羊男》如果是英語片，當年會得奧斯卡 2/08 tontontonni
8 Re: [討論]環太平洋2 中國神翻譯預告 2/08 shirman
[討論]靈魂交換新片 2/08 nancy3nancy3
16 [問片]一部古埃及詛咒的老恐怖片神鬼艷后/天煞 2/08 NaaL
9 Re: [贈票]《BPM》台北特映會搶先看 2/08 xup6m3rmp
2 (本文已被刪除) 2/08 -
Re: [討論]環太平洋2 中國神翻譯預告 2/08 cloud72426
1 [討論]縮小人生一幕 2/08 kc092444
11 [討論]像"艾蜜莉布朗"一樣又正又魯的女星? 2/08 peter080808
49 [新聞]《猛毒》前導預告釋出 2/08 rz759
3 [問片]小女孩是惡鬼/惡魔？ 2/08 suumire
1 [普好雷]《誰是大壞狐》 2/08 janice100
3 [極好雷] 縮小人生：存在主義治療 2/08 gushing
X1 (本文已被刪除) 2/08 -
7 [討論]霓裳魅影的Cyril 2/08 dragon50119
11 [情報]艾米漢默將主演《闇影之下》導演編導未命名新片 2/08 qpr322
1 [問片]一個人被卡在工地電梯上 2/08 ting19841207
3 Re: [贈票]票選2017的年度電影片單(贈票公布＆候補) 2/08 popchieh
2 [普雷]格雷的五十道陰影-自由 2/09 babe18
0 0 Python 3.6 (64-bit)
第 2 行，第 30 欄 空格: 4 UTF-8 CRLF Python

project01.py - py - Visual Studio Code

檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)

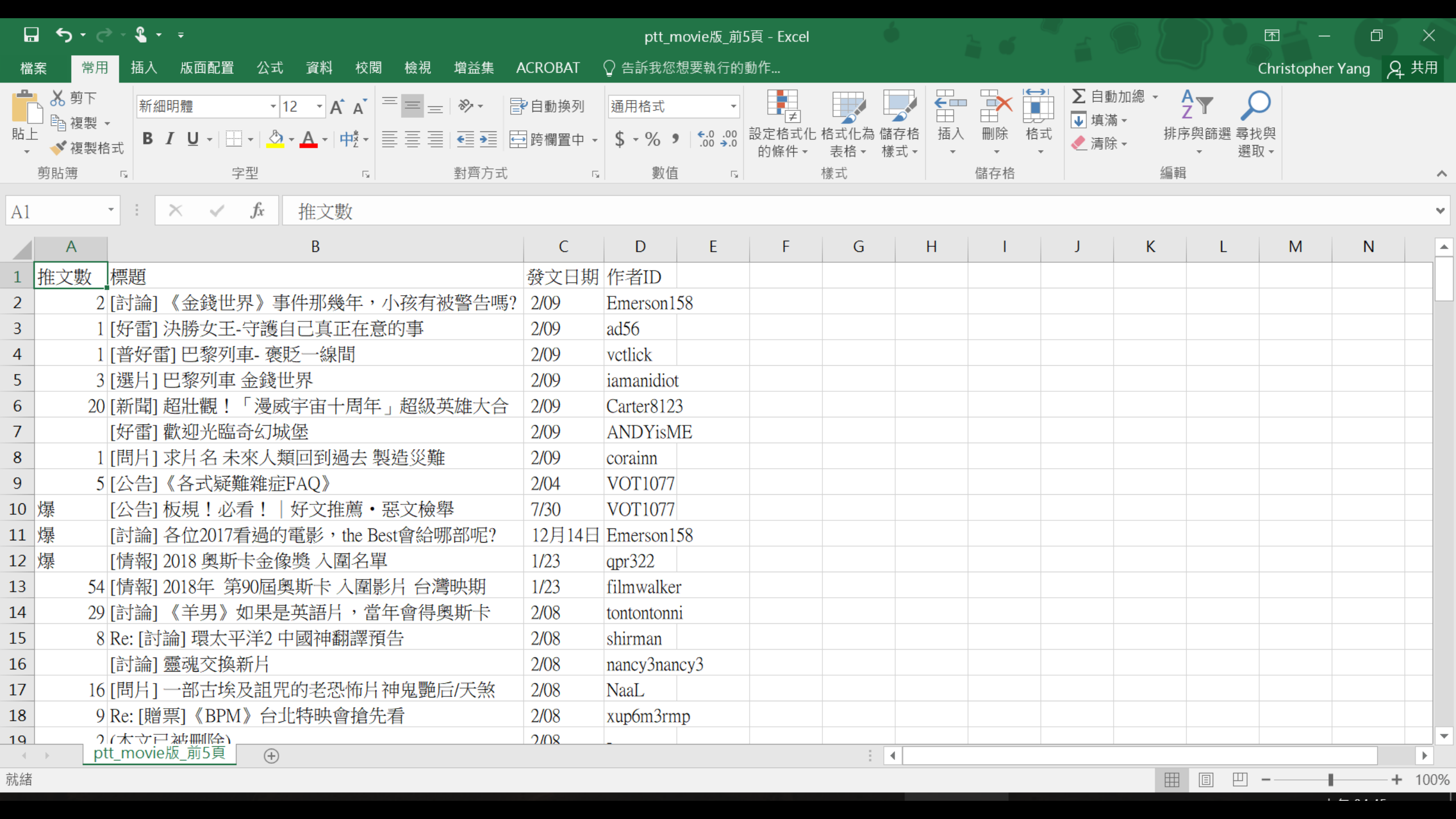
project01.py

```
1  import requests
2  from bs4 import BeautifulSoup
```

問題 輸出 偵錯主控台 終端機 2: Python

```
 3    [問片] 『港片』全家參加省錢比賽                    2/08 xinghh
      [新聞] 《格雷3》巴黎首映 「總裁夫人」美穿PRAD      2/08 huanglove
24    [新聞] 詹姆斯法蘭柯 衝浪避鋒頭揪爽                2/08 huanglove
 8    [請益] 雷-霓裳魅影 衣服的內襯                     2/08 filmmaker
31    [新聞] 63歲男星蜜戀31歲小模  拖了2年終於要離       2/08 huanglove
 6    (本文已被刪除)                                   2/08 -
 5    [好雷] 誠意滿點之角頭2                           2/08 akbobo
42    [新聞] 《逃出絕命鎮》拍續集？  導演：有很多故     2/08 sosoing
 4    Re: [情報] 第37屆香港電影金像獎提名名單          2/08 tools
 2    [好雷] 王牌業務員Es war einmal in Deutschland    2/08 mysmalllamb
 2    [新聞] 韓國女導演李賢珠被曝性侵女同事            2/07 hahaha0204
      [負雷] 決勝女王  覺得編劇不專業                   2/07 inbow
 3    [討論] 池畔謎情看完的疑惑                        2/07 s2657507
38    [選片] 花甲大人轉男孩  格雷的五十道陰影           2/07 babe18
      Re: [新聞]NETFLIX以超過五千萬美元取得《科洛弗悖論 2/07 femlro
50    [討論]「死侍2」最新預告                          2/07 beckseaton
      (本文已被刪除)                                   2/07 -
22    [好雷] 水底情深 - 我的小怪物你在哪？             2/07 priliona
      [贈票]《獄火重生:金昌洙》台北台中特映會          2/07 g3yu
 3    [好雷] 《霓裳魅影》 - 不自知的虐戀渴望           2/07 leila
10    [普雷] 《科洛弗悖論》-行銷手法的驚喜大過電影     2/07 leila
11    (本文已被刪除)                                   2/07 -
 2    [新聞] 「小美」、「十四顆蘋果」 出征柏林影展     2/07 iam168888888
 爆   Re: [新聞] 【不可能的任務：全面瓦解】最新預告   2/07 draft
 6    [好無雷] 角頭2 國產黑幫片誠意之作                2/07 tieamonk
 2    [片單] 類似金牌黑幫的電影                        2/07 HOWARDNO1
24    [請益] 第一次看格雷可以懂劇情嗎？                2/07 syensyens
 1    [新聞] 特派專欄：寶萊塢電影引發暴動背後的真相    2/08 sarada
17    [討論] 金馬獎的口味是不是很難捉摸?               2/08 takuminauki
 2    Re: [討論] <氣象戰>裡的一個疑問                   2/08 imrt
輸入 y 以匯出成csv檔，輸入其他結束程式：
```

0 0 Python 3.6 (64-bit) 第 2 行，第 30 欄 空格: 4 UTF-8 CRLF Python



推文數	標題	發文日期	作者ID
2	[討論]《金錢世界》事件那幾年，小孩有被警告嗎?	2/09	Emerson158
1	[好雷] 決勝女王-守護自己真正在意的事	2/09	ad56
1	[普好雷] 巴黎列車- 褒貶一線間	2/09	vctlick
3	[選片] 巴黎列車 金錢世界	2/09	iamanidiot
20	[新聞] 超壯觀！「漫威宇宙十周年」超級英雄大合	2/09	Carter8123
	[好雷] 歡迎光臨奇幻城堡	2/09	ANDYisME
1	[問片] 求片名 未來人類回到過去 製造災難	2/09	corainn
5	[公告]《各式疑難雜症FAQ》	2/04	VOT1077
爆	[公告] 板規！必看！｜好文推薦・惡文檢舉	7/30	VOT1077
爆	[討論] 各位2017看過的電影，the Best會給哪部呢?	12月14日	Emerson158
爆	[情報] 2018 奧斯卡金像獎 入圍名單	1/23	qpr322
54	[情報] 2018年 第90屆奧斯卡 入圍影片 台灣映期	1/23	filmwalker
29	[討論]《羊男》如果是英語片，當年會得奧斯卡	2/08	tontontonni
8	Re: [討論] 環太平洋2 中國神翻譯預告	2/08	shirman
	[討論] 靈魂交換新片	2/08	nancy3nancy3
16	[問片] 一部古埃及詛咒的老恐怖片神鬼艷后/天煞	2/08	NaaL
9	Re: [贈票]《BPM》台北特映會搶先看	2/08	xup6m3rmp
2	(本文已被刪除)	2/08	

X1 [請益] 請問有人不用蘋果手機,微軟辦公,谷歌搜尋 2/08 candybeer

Thank You